Journal of Informetrics 8 (2014) 569–580

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Including cited non-source items in a large-scale map of science: What difference does it make?

Kevin W. Boyack[a,*], Richard Klavans[b]

[a] SciTech Strategies, Inc., Albuquerque, NM 87122, USA
[b] SciTech Strategies, Inc., Berwyn, PA 19312, USA

ARTICLE INFO

Article history:
Received 3 February 2014
Received in revised form 28 March 2014
Accepted 2 April 2014
Available online 17 May 2014

Keywords:
Science mapping
Direct citation
Non-source documents
Books

ABSTRACT

Cited non-source documents such as articles from regional journals, conference papers, books and book chapters, working papers and reports have begun to attract more attention in the literature. Most of this attention has been directed at understanding the effects of including non-source items in research evaluation. In contrast, little work has been done to examine the effects of including non-source items on science maps and on the structure of science as reflected by those maps. In this study we compare two direct citation maps of a 16-year set of Scopus documents – one that includes only source documents, and one that includes non-source documents along with the source documents. In addition to more than doubling the contents of the map, from 19 M to 43 M documents, the inclusion of non-source items strongly augments the social sciences relative to the natural sciences and medicine and makes their position in the map more central. Books are also found to play a significant role in the map, and are much more highly cited on average than articles.

1. Introduction

One of the goals of science mapping, whether mapping full databases or smaller local datasets, has been to map the associated topic space as accurately as possible. It is well known, however, that coverage of the scholarly literature in the dominant databases (e.g., Web of Science, Scopus, PubMed) varies widely by discipline. Coverage is typically high in natural science disciplines such as chemistry and physics, slightly lower in the medical sciences, lower still in technical fields such as engineering, and very low in the social sciences and humanities (Butler & Visser, 2006; Hicks, 2004; Moed, 2005; van Leeuwen, 2006). Thus, we can assume that while maps of scientific areas in chemistry and physics will have close to full coverage of the topic space, maps in other disciplines may miss some topics altogether simply because of lack of database coverage of the literature associated with those topics. Global mapping, or mapping of all of science, may be especially vulnerable to the effects of this variance in coverage by discipline because all disciplines are present in a single map.

Although most science mapping efforts to date have focused on what are referred to as source items (publications in sources indexed by the database provider), there are no inherent limitations in science mapping techniques that would preclude non-source items from being included in a map. Any item in the data, whether source or cited non-source, can be mapped provided there is some information about that item which links it to other items. For example, title words can

* Corresponding author. Tel.: +1 505 856 1267.
E-mail addresses: kboyack@mapofscience.com (K.W. Boyack), rklavans@mapofscience.com (R. Klavans).

http://dx.doi.org/10.1016/j.joi.2014.04.001
1751-1577/

be used to include non-source items in text-based maps, and citation links from source to non-source items can be used to include non-source items in citation-based maps.

In this study we characterize the effects of including large numbers of non-source items in a global map of science. Two maps generated from the same set of database records and using a similar mapping methodology are compared. One map includes only source items; the second map includes source items and those non-source items that are cited at least twice. The balance of this paper proceeds as follows. First, relevant literature is reviewed. Data and methods used are then described, followed by a characterization of the two maps. Significant differences between the maps and the literature they represent are presented. The paper concludes with a discussion of the implications of this work on the characterization of science and technology.

2. Background

2.1. Source vs. non-source items

References cited by source documents in citation databases can be divided into two types:

- source items – references for which an indexed source record exists in the database,
- non-source items – references for which an indexed source record does not exist in the database.

While source items typically comprise around 75% of all references for a single publication year in the Web of Science, these numbers vary dramatically by discipline, ranging from around 90% for molecular biology and chemistry to less than 20% for the humanities (Moed, 2005). Although exact numbers vary, other studies show similar fractions of non-source items for the same broad areas of science (Butler & Visser, 2006; van Leeuwen, 2006). Hicks (2004) shows that while 85% of the output from natural scientists is in the form of journal and conference papers, the number is only around 50% for social scientists. An earlier study by Hicks (1999) reports that books comprise between 40% and 60% of the social science literature.

Non-source items are known to consist of many different document types. These include journal articles from non-indexed sources, conference papers, books, handbooks, book chapters, monographs, working papers, corporate and government reports, software, and even articles from newspapers such as the *New York Times*. Of these many document types, books seem to be getting the most recent attention. Nederhof, van Leeuwen, and van Raan (2010) analyzed highly cited non-source items in psychology and political science, finding that for references published after 1980, books formed the majority of these highly cited non-source items. Huang and Chang (2008) surveyed previous studies showing that books comprised from 15% to 89% of cited sources in various fields in the social sciences and humanities; books comprised more than half of all cited sources in 17 of the 25 individual cases surveyed. Zuccala and Guns (2013) classified documents cited by articles in over 1000 humanities journals and found there were more citations to books than to other document types combined.

More work has been done to characterize the effects of including (or not including) non-source material on research evaluation than upon science mapping, particularly in the social sciences and humanities where citations to non-source items such as books are known to be prevalent. For example, Butler and Visser (2006) performed an extensive bibliometric analysis of non-source items published by Australian universities, finding that they can substantially augment publication and citation counts in the social sciences and humanities, and can have a significant effect on rankings. Nederhof (2006) reviews efforts to address research performance in social sciences and humanities using bibliometrics and concludes that non-source items need to be included. More recently, Chi (2013) found that the inclusion of non-source items in evaluation of political science researchers significantly increases the numbers of publications reported, but has a much milder effect on their H-index values. We note that Google Scholar is gaining traction as a source for such evaluations given that non-source items seem to be extensively covered (Franceschini & Maisano, 2011).

2.2. Mapping of non-source items

From their earliest days, science mapping efforts have routinely included non-source items. In fact, non-source items were far more prevalent in early science maps than they are today. The earliest common implementations of direct citation maps, Garfield's historiographies (Garfield, 1973), did not distinguish between source and non-source items. This was also true for early document co-citation (Small, 1973) and author co-citation (White & Griffith, 1981) maps. These early studies simply mapped documents or authors, and paid no attention to the distinction between source and non-source items. The way in which citation indexes evolved played a role in this. In the 1970s and 1980s, data for many science maps was extracted from print editions of the (Social) Science Citation Index, or from electronic compilations of these data in DIALOG. These sources included lists of cited items, enabling datasets and maps to be created based on cited documents and authors, many of which did not appear as source items in the data. As the citation indexes moved from print to CDROM versions, and finally to the fully searchable Internet-based platforms of today, datasets for mapping have increasingly been constructed based on searches of source items.

Mapping of non-source journals has rarely been done. Tijssen and van Leeuwen (1995) mapped a combined set of source and non-source journals in the area of manufacturing technology and management. This required merging of data from three sources – JCR, Compendex and Ulrich's International Dictionary of Periodicals. A source journal map based on the JCR

Table 1
Characterization of Scopus data.

Year	# Records	# Records with references	% Records with references	# References	% References to 1996–2011 source items
1996	1,134,758	785,196	69.2	20,874,374	1.08
1997	1,161,780	813,750	70.0	21,719,692	6.02
1998	1,164,390	828,917	71.2	22,541,903	12.80
1999	1,166,048	849,141	72.8	23,782,665	19.76
2000	1,224,001	919,587	75.1	25,837,244	25.98
2001	1,325,284	1,009,738	76.2	27,518,645	31.70
2002	1,374,293	1,052,326	76.6	29,362,784	36.46
2003	1,429,751	1,125,557	78.7	31,564,418	40.64
2004	1,578,957	1,261,066	79.9	34,736,118	44.20
2005	1,755,980	1,394,297	79.4	38,605,815	47.08
2006	1,843,420	1,519,643	82.4	42,426,261	49.59
2007	1,944,239	1,630,369	83.9	46,085,026	50.82
2008	2,020,576	1,724,663	85.4	49,439,868	52.88
2009	2,110,248	1,865,368	88.4	54,088,145	54.92
2010	2,219,650	1,969,807	88.7	58,133,563	57.11
2011	2,352,087	2,074,973	88.2	62,374,997	59.38
Total	25,805,462	20,824,398	80.7	589,091,518	42.81

only was compared with a combined map based on content information (journal descriptors). Experts found the content-based map to be much more comprehensive and accurate than the citation-based map. More recently, Leydesdorff (2008) mapped the citation impact environment of *Science and Public Policy*, a non-source journal at the time, using cited references from source journals. van Eck and Waltman (2010) used the VOSviewer system to generate a journal co-citation map from references cited by documents published in 2007 from WoS data. Although no differentiation was made between source and non-source journals in the cited references, non-source journals were not highlighted, so their prevalence in the map is unknown.

Some recent document-level science mapping studies have included non-source materials. For example, Chen and colleagues routinely include highly cited non-source items in their document co-citation analyses and maps (Chen, 2006; Chen, Ibekwe-SanJuan, & Hou, 2010; Chen & Kuljis, 2003). Noyons and Calero-Medina (2009) included noun phrases parsed from the titles of non-source papers in their maps of the research areas of three Dutch Universities of Technology. However, despite studies like these, non-source items have not been included systematically in any large-scale map of science. This study is the first to include large numbers of non-source items in a map of all of science.

3. Data and methods

A 16-year (1996–2011) set of Scopus data was used to create our two maps of science. These data were obtained from Elsevier in summer 2012, and thus contained a nearly complete set of 2011 data. The data from those years is comprised of 25.8 million records of which 20.8 million have references (see Table 1). There are also a total of 589 million references (citing-cited pairs) to 115 million unique cited documents, 34.8 million of which were cited at least twice. Scopus document IDs were used in this study; no additional work was done to clean or unify cases where multiple document IDs refer to the same document.

Given that Scopus only indexes references from items published in 1996 and later, non-source items consist of all references published prior to 1996 as well as those references published later than 1995 that are not source items. Thus, the fraction of references to source items in 1996 is very small, and increases with each subsequent year. We note that the fraction of source items reaches a maximum of 59.4% in 2011, but never reaches the 75% value quoted by Moed (2005). This is likely due to the fact that we only have 16 years (rather than 20+) of source items that can be matched, and also that Scopus is enriched in social science and humanities sources, which have lower reference rates to source items than the natural sciences (Leydesdorff, 2003).

Two separate maps were created from this set of documents and references. For the first map (referred to hereafter as SRC), the citing-cited pairs were limited to cited documents that were also source items. Thus, all references used to create this map were between pairs of source items, a total of 252.2 million citing-cited pairs. For the second map (referred to hereafter as NS), citations to non-source items that were cited at least twice were included along with the citations to source items, comprising a total of 510.7 million citing-cited pairs. Of cited non-source items, those cited only once are the least influential (as measured by citation counts), and in many cases may be random events. Thus, the approximately 80 million non-source items cited only once were excluded from the process.

The process used to create the SRC map is detailed here. First, similarity values are calculated for each pair of documents linked by direct citation using the citing-cited pairs. Second, documents are clustered using the CWTS modularity-based code of Waltman and van Eck (2012). Third, a visual layout of the clusters is created using textual similarity between clusters and the DrL graph layout algorithm. Each of these steps is given in more detail below.

(1) Similarity between pairs of documents based on direct citation was calculated as follows:
- Coefficients a_{ij} are calculated as $a_{ij} = 1.0/n_{citing}$, where n_{citing} is the number of papers within the set that are referenced by the citing paper in pair *i,j*. We did not see the need to account for times cited in addition to the number of references in our normalization scheme because each citing paper contributes a total weight of 1.0 to the system.
- We do not use the raw a_{ij} values as similarities, but take an additional step. Since, the a_{ij} values comprise only one half of a full matrix, we set $a_{ji} = a_{ij}$ to form a symmetric matrix. K50 (essentially a cosine minus its expected value) coefficients are then calculated as

$$K50_{ij} = K50_{ji} = \max\left[\frac{(a_{ij} - E_{ij})}{\sqrt{S_i S_j}}, \frac{(a_{ji} - E_{ji})}{\sqrt{S_i S_j}}\right]$$

where $E_{i,j} = (S_i S_j)/(SS - S_i)$, $S_i = \sum_{j=1}^{n} F_{i,j}$, $\quad j \neq i$, $SS = \sum_{i=1}^{n} S_i$

E is an expected value of *a*, and varies with S_j.
- To reduce the number of similarity values that are input to the CWTS clustering routine, we filtered the K50 values using our typical approach, which keeps only the top-N most related documents (highest similarities) for each document. The total degree (inlinks + outlinks) for each document is calculated, and the range encompassed by the resulting degree distribution is scaled using log(degree) values to a 5–15 scale. The degree for each document thus determines how many pairs that document brings into the final similarity file, varying between 5 and 15 similarity pairs per document. After top-N filtering the number of similarity pairs used in the SRC calculation was 91,708,923.

(2) The new modularity-based code from CWTS was used to cluster the direct citation similarity file from step (1). Although the CWTS code is capable of generating a hierarchical model with nested clusters, we chose to run it at a single level which is roughly equivalent to level 3 in the original calculation of Waltman and van Eck (2012). The SRC calculation was run with a minimum cluster size (n_{min}) of 20, and resolution (r) of 9.0×10^{-5}. The resolution value was chosen such that the solution would give on the order of 150,000 clusters. The code was run 10 times with different random restart values and the solution that maximized the CWTS quality function was used as our completed model.

(3) A layout and visual map of the clusters was created using textual similarity. Textual similarity is used in place of citation-based similarity in this step because it has been found to give maps that are more visually appealing and more accurate from an author consistency point of view (Boyack & Klavans, 2014). BM25 similarity values between pairs of clusters were computed for all pairs of clusters using the titles and abstracts of document in the clusters. The BM25 similarity between one object *q* and another object *d* is calculated as:

$$s(q,d) = \sum_{i=1}^{n}\left(\mathrm{IDF}_i \frac{n_i(k_1+1)}{n_i + k_1((1-b+b^{|D|})/\bar{D})}\right)$$

where n_i is the frequency of term *i* in object *d*. Note that $n_i = 0$ for terms that are in *q* but not in *d*. Typical values were chosen for the constants k_1 and *b* (2.0 and 0.75, respectively). In our formulation each cluster was treated as if it were a single document. Document length $|D|$ was estimated by adding the term frequencies n_i per document. Average document length $|\bar{D}|$ is computed over the entire set of documents. The IDF value for a particular term *i* is computed as:

$$\mathrm{IDF}_i = \log\frac{N - n_i + 0.5}{n_i + 0.5}$$

where *N* is the total number of documents in the dataset and n_i is the number of documents containing term *i*. Each individual term in the summation in the first formula is independent of document *q*. To remove the influence of high frequency terms all IDF scores below 2.0 were discarded.

A matrix of text-based similarity values is typically far less sparse than a matrix composed of citation-based similarities. Thus, once BM25 values were calculated, we once again filtered the similarities to the top-N similarity values per cluster, using the same filtering method mentioned above. The resulting similarity files were input to the DrL (now known as OpenOrd) graph layout routine (Martin, Brown, Klavans, & Boyack, 2011), which calculates an *x,y* position for each input object based on the similarity values (weighted edges) that were input. The resulting map was rotated and/or flipped to produce an orientation that is consistent with our previous maps, with physics at the top of the map and chemistry at the right-hand side.

The process that was used to create the NS map was identical to that used for the SRC map with two exceptions. First, the top-*N* filtering step to reduce the number of similarity values was not run for the NS map. Thus, the full set of 510,696,975 pairs was used in this calculation. Note that this NS clustering calculation was run on an Amazon EC2 server with 64 GB memory; each run required roughly 20 h of CPU time. With server costs at less than $2/h, the set of 10 runs cost around $400 to run on the Amazon server. Second, the NS clustering calculations were run with a minimum cluster size (n_{min}) of 50, and

Table 2
Document distributions for the SRC and NS cluster solutions.

Year	# Records with references or cites	SRC map		NS map			
		# Records	%Coverage	# Records	# Source items	%Coverage	# Non-source items
Pre-1961		53[a]		1,213,097			1,213,097
1961–1970				1,047,503			1,047,503
1971–1980				2,357,758			2,357,758
1981–1985				2,097,808			2,097,808
1986–1990				3,217,162			3,217,162
1991–1995		342[a]		5,094,104			5,094,104
1996	932,810	669,735	71.8	1,420,077	880,324	94.4	539,753
1997	955,309	712,141	74.5	1,445,240	903,194	94.5	542,046
1998	967,968	745,083	77.0	1,476,665	918,212	94.9	558,453
1999	981,966	778,341	79.3	1,491,882	933,069	95.0	558,813
2000	1,037,420	847,454	81.7	1,581,096	991,758	95.6	589,338
2001	1,093,789	933,579	85.4	1,618,682	1,053,081	96.3	565,601
2002	1,141,197	969,394	84.9	1,674,353	1,099,162	96.3	575,191
2003	1,209,760	1,044,161	86.3	1,736,076	1,166,742	96.4	569,334
2004	1,347,975	1,172,113	87.0	1,833,955	1,300,578	96.5	533,377
2005	1,479,417	1,288,923	87.1	1,913,140	1,426,956	96.5	486,184
2006	1,594,556	1,406,705	88.2	1,980,664	1,541,889	96.7	438,775
2007	1,695,733	1,502,622	88.6	2,013,494	1,639,098	96.7	374,396
2008	1,791,193	1,596,413	89.1	2,037,512	1,732,201	96.7	305,311
2009	1,906,214	1,693,916	88.9	2,071,144	1,852,002	97.2	219,142
2010	2,000,426	1,738,522	86.9	2,052,318	1,947,626	97.4	104,692
2011	2,083,850	1,913,081	91.8	2,057,858	2,038,437	97.8	19,421
Total	22,219,583	19,012,578	85.6	43,431,588	21,424,329	96.4	22,007,259

[a] Non-zero numbers here due to mismatches in file year and publication year.

resolution (r) of 0.0975. Once again, the resolution value was set such that the solution would have around 150,000 clusters, thus making it easier to compare the results of the SRC and NS maps.

4. Results and discussion

4.1. Clustering

The two clustering calculations gave results with similar numbers of clusters. The SRC map is comprised of 19,012,578 source documents in 149,613 clusters, while the NS map is comprised of 21,424,329 source and 22,007,259 non-source documents in 156,085 clusters. Document distributions by year are given in Table 2. Publication years for the non-source items are known because this information is included (along with titles, sources, and most authors) in the XML reference data for the source items.

The second column of Table 2 lists the numbers of Scopus records per year that have at least one reference or have been cited at least once. Since only those documents with links can be included in a direct citation map, this represents the maximum number of source documents that can appear in a map, and is an appropriate number with which to compute coverage. The SRC map includes 85.6% of all possible source documents, with coverage varying by year. Early years are less represented than later years. This is natural because the within-set links for documents in early years are dominated by inlinks (being cited) rather than outlinks (references). In most cases an early paper will only be included in the map if it has been cited. Only 71.9% of Scopus source items from 1996 to 2001 have been cited by 2011; this is the minimum coverage that a direct citation map should achieve. The SRC coverage percentages listed in Table 2 suggest that most early papers are indeed linked into the map by citations from subsequent papers, and that any coverage above this 71.9% value is due to additional papers having been linked into the map through their references.

Comparison of the coverage of the SRC and NS cluster distributions leads to some fascinating observations. First, coverage of source items increases dramatically in the NS map, to 96.4%. This happens because cited non-source items can link source papers into the map, regardless of when the non-source items were published. Given that the top-N filtering of links was not applied to this NS map, one might expect full (100%) coverage of source items. However, we excluded ~80 M non-source items that were cited only once. Any source papers whose only links into the map would have been through these excluded non-source items are not included in the map, thus leading to coverage values of less than 100%.

More cited non-source items than source items appear in the NS map. This is to some degree an effect of the fact that Scopus does not include source items prior to 1996. However, there are significant numbers of non-source items (nearly 7 million) from 1996 to 2011 that appear in the map, augmenting the content of the map by 1/3 in those source years. This additional non-source content has the potential to significantly affect the distribution of documents by field and our perceptions of the structure of science.

Table 3
Document distributions by major field for the SRC and NS maps.

Major field	SRC map			NS map			% Chg
	# Clusters	# Docs	% Docs	# Clusters	# Docs	% Docs	(NS-SRC)
Comp Sci/EE	17,231	2,229,768	11.73	17,530	4,569,978	10.52	−1.21
Math/Physics	13,088	2,077,607	10.93	10,350	4,198,377	9.67	−1.26
Chemistry	19,004	2,141,329	11.26	15,986	4,531,179	10.43	−0.83
Engineering	13,631	1,922,512	10.11	18,821	4,676,011	10.77	+0.65
Earth Sciences	3936	543,746	2.86	7607	1,798,168	4.14	+1.28
Biology/Biotech	12,439	1,620,208	8.52	15,446	4,299,023	9.90	+1.38
Infectious Disease	6184	817,736	4.30	4890	1,753,092	4.04	−0.26
Medical Specialties	31,835	3,924,438	20.64	17,125	7,379,251	16.99	−3.65
Health Sciences	10,160	1,265,745	6.66	11,362	2,974,390	6.85	+0.19
Brain Research	9266	1,122,491	5.90	5462	2,247,977	5.18	−0.73
Social Sciences	12,334	1,298,295	6.83	27,800	4,561,768	10.50	+3.67
Humanities	480	47,554	0.25	2748	351,571	0.81	+0.56
Not classified	25	1149	0.01	958	90,803	0.21	+0.20
Total	149,613	19,012,578		156,085	43,431,588		

4.2. SRC and NS maps

Visual maps have been created from the two cluster solutions (SRC and NS) using the text-based layout method described above. Most source articles in each map were assigned colors based on the color-to-journal-to-major field scheme used in the UCSD map of science (Börner et al., 2012), and each cluster was colored by dominant article color. Each of the 12 colors used in the maps represents a major field (e.g., Chemistry, Engineering, Biology, Social Sciences). This allows us to examine each map of science and to compare document distributions by major field. Gray was used to color the clusters that could not be classified using source paper colors (e.g., at the lower left of the NS map).

Fig. 1 shows that the upper halves of the SRC and NS maps, above the line stretching roughly from the 10:00-to-4:00 positions (using a clock metaphor) are quite similar. The relative positions of the groups of clusters in Computer Science (pink), Math/Physics (purple), Chemistry (blue), Engineering (cyan), Earth Sciences (brown), and Biology (green) are roughly the same. There are, of course, local differences, but the high level structure seen in the SRC map is preserved in the NS map. In contrast, the lower halves of the two maps show more differences. The adjacencies of major fields (colors that are next to each other) are the same in both maps. However, the relative sizes of the fields and their absolute positions have changed. The Social Sciences (light orange) appears as a single set of connected islands at the far left of the SRC map. However, in the NS map the Social Sciences form several disconnected islands which together take up far more space than is occupied by the Social Sciences in the SRC map. Furthermore, the main Social Sciences island appears at the interior of the map rather than at the edge, and seems to have pushed the medical fields (reds and yellow) toward the bottom of the NS map. One interesting aspect of this large Social Sciences island is that it seems to be surrounded by far more white space than any other island in the map. It is as if this island is largely self-contained and is keeping other fields at a distance. There are but a few small trails of clusters between this island and any of the other areas in the map. This picture of the Social Sciences is consistent with the findings of Bollen et al. (2009), whose map of science based on clickstream log data showed the Social Sciences at the center of the map, but with few links to the natural sciences and medicine.

The effect of the inclusion of non-source items on the distribution of documents by major field has also been investigated. Numbers of documents by major field were estimated for both maps by summing the numbers of documents in clusters for each color. This assumes that all documents in a cluster belong to the same major field. This assumption is very reasonable from the perspective that all documents in a cluster are related to a single topic through their direct citation links, and thus are part of the major field assigned to the cluster, regardless of which journal they appeared in. Table 3 shows that inclusion of non-source materials does shift the distribution of documents by major field in a significant way. The fraction of documents in Medical Specialties decreases by 3.7% while the fraction of documents in the Social Sciences increases by 3.7%. Humanities sees a three-fold increase, from 0.25% of documents to 0.81% of documents. Decreases are also seen in Computer Science, Math/Physics, and Chemistry and Brain Research, while increases are seen in Engineering and the Earth and Biological Sciences. Although nothing definitive can be said about the clusters that could not be classified by major field (those that are gray), their position in the NS map suggests that most are likely associated with the Social Sciences and Humanities.

It is also interesting to look at the field-wise distributions of the NS map with respect to source and non-source items and different time periods. The 43.4 million documents in the NS map can be divided into five groups:

- Src9611: Source items, all of which were published from 1996 to 2011
- Nonsrc9611: Non-source items published from 1996 to 2011

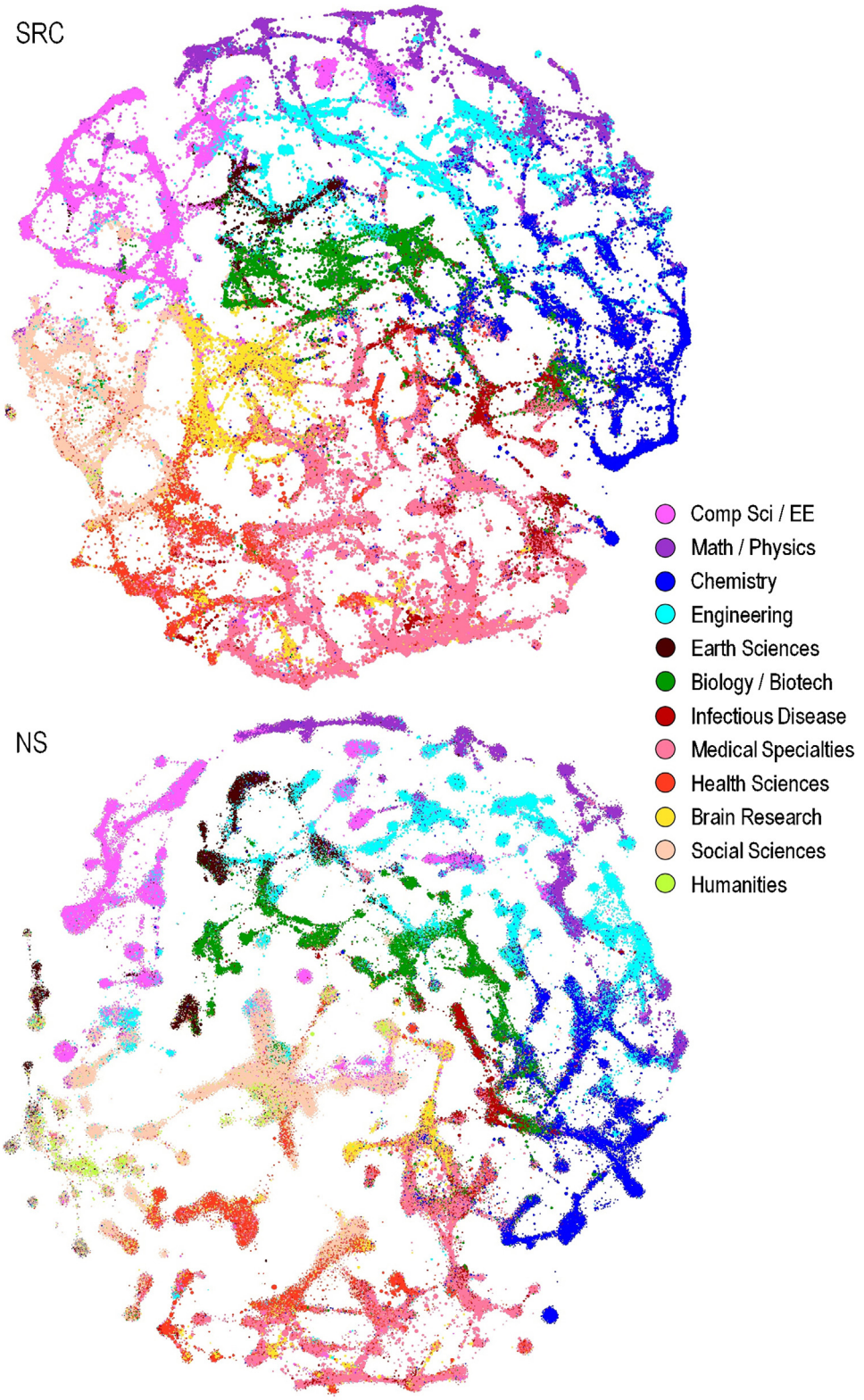

Fig. 1. Maps of science based on source-only (SRC) and source + non-source items (NS). High-resolution maps are available at http://www.mapofscience.com/?page_id=790. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

Table 4
Source and non-source items by major field and time period for the NS map.

Major field	Src9611		Nonsrc9611		SC6195		Nonsrc6195	
	# Docs	% Docs	# Docs	% Docs	# Docs	% Docs	# Docs	% Docs
Comp Sci/EE	2,584,541	12.06	1,242,536	17.81	275,898	3.95	394,790	5.78
Math/Physics	2,264,909	10.57	495,660	7.11	639,507	9.16	683,791	10.00
Chemistry	2,183,115	10.19	467,426	6.70	913,089	13.08	838,422	12.26
Engineering	2,307,625	10.77	902,506	12.94	561,876	8.05	803,278	11.75
Earth Sciences	621,688	2.90	347,531	4.98	221,333	3.17	508,197	7.43
Biology/Biotech	1,721,288	8.03	612,828	8.79	669,060	9.58	1,077,729	15.77
Infectious Disease	903,498	4.22	163,179	2.34	447,455	6.41	207,323	3.03
Medical Specialties	4,387,804	20.48	584,941	8.39	1,868,558	26.76	430,827	6.30
Health Sciences	1,521,527	7.10	442,704	6.35	596,147	8.54	348,964	5.10
Brain Research	1,219,909	5.69	213,506	3.06	554,848	7.95	212,914	3.11
Social Sciences	1,605,928	7.50	1,400,150	20.07	226,325	3.24	1,171,969	17.14
Humanities	85,089	0.40	82,441	1.18	7887	0.11	121,546	1.78
Not classified	17,408	0.08	19,999	0.29	628	0.01	36,394	0.53
Total	21,424,329		6,975,407		6,982,611		6,836,144	

- SC6195: Items contained in the Scopus pre-1996 archive set.[1] Scopus has added, and continues to add, large backfile collections from major publishers. There are currently 21 million of these archived records which differ from the Src9611 source items in that they do not contain references, and thus can only appear in the map when referenced by Src9611 items. Thus, although they are sourced by Scopus, for purposes of this analysis we consider these as non-source items.
- Nonsrc6195: Non-source items published from 1961-1995 that are not present in the SC6195 archive set.
- Other: Non-source items published prior to 1961 or which have obviously erroneous publication years (e.g., 3000).

Table 4, which gives numbers of documents by major field for the first four of these groups, shows that the distributions of source and non-source items by major field are quite different. When considering the 1996–2011 time period, the medical fields (Infectious Disease, Medical Specialties, Health Sciences, Brain Research) comprise 37.5% of the source items, but only 20.1% of the non-source items. In contrast, while Computer Science, Engineering, and the Social Sciences comprise only 30.3% of the source items, more than half (50.8%) of the non-source items come from these fields. Earth Science and Humanities also have higher fractions of non-source than source items, while Physics and Chemistry have lower fractions of non-source than source items. Similar patterns are seen for the 1961–1995 time period. The difference between fractions of source (49.7%) and non-source (17.6%) items for medical fields is even more pronounced. The Social Sciences are once again the greatest beneficiaries of including non-source items and have the largest fraction of non-source items during this time period.

Differences in the distributions between time periods are also instructive. For example, the footprint of Computer Science is much smaller in the earlier time period than it was in the later time period. This likely reflects both the recent growth in that field and the fact that the Scopus archive set likely contains few computer science source titles and no early proceedings. In contrast, Chemistry comprises a larger fraction of the whole in the earlier time period than in the later time period. This likely reflects the fact that Chemistry was already a mature field in the earlier time period. Surprisingly (to us, at least), Biology had a much larger fraction of the non-source items from 1961 to 1995 (15.8%) than it had in the 1996–2011 time period (8.8%). This suggests that not only was Biology a mature field in the earlier time period, but also that much of its important literature was not being indexed at that time. Many additional observations are possible from the data in Table 4. On the whole, the differences between the source and non-source distributions correlate very well with previous investigations of cited non-source items (Butler & Visser, 2006; Hicks, 2004; Moed, 2005). Non-source fractions are higher than source fractions in fields where the majority of references are to non-source items.

4.3. Non-source document types

Given current interest in the impact of non-source items, we also attempted to identify items of different types to determine the effect of different document types on the NS map. Accurate identification of document type (e.g., journal article, book) is not possible for all non-source items. Nevertheless, some heuristics are available which allow us to make educated guesses of type for many items. For example, Nederhof et al. (2010) suggested that journal articles typically have volume and page numbers, while other document types typically have neither. We found this to be true for the most part, but also found that many books and handbooks have values in the volume and page fields, often representing edition numbers, etc. Regarding books, it was suggested to us that books could be identified as those items where source (e.g., journal name) and title strings are identical (personal communication associated with Zuccala & Guns, 2013). Upon spot-checking numerous examples, we found this to be true, and to be a very useful way of identifying books. It is not clear that all books have this

[1] http://cdn.elsevier.com/assets/pdf_file/0019/148402/contentcoverageguide-jan-2013.pdf, page 22.

Table 5
Document type counts and impacts for four document groups in the NS map.

Document type	Src9611		Nonsrc9611		SC6195		Nonsrc6195	
	# Docs	Avg cites	# Docs	Avg cites	# Docs	Avg cites	# Docs	Avg cites
Jnl/conf paper	21,417,462	11.75	3,335,565	5.36	6,905,566	20.13	3,725,537	9.59
Book	1136	51.64	303,834	28.10	4966	48.36	188,800	69.69
Handbook	5731	6.87	65,215	9.90	946	26.69	46,604	12.86
Source/no title			1,807,465	4.34			1,820,089	5.37
Not classified			1,463,328	5.03	71,133	12.80	1,055,114	6.78
Total	21,424,329		6,975,407		6,982,611		6,836,144	

profile. Thus we also spot-checked items with a source string but no title string. Although some of these items are books, we also found other document types, such as reports and chapters, in this set of documents.

Ultimately, we separated documents into five types using the following process:

- Books: documents where source = title.
- Handbooks: documents with the word 'Handbook' in the source string.
- Jnl/Conf papers: all 1996–2011 source documents that were not books or handbooks, and all other documents with volume and page numbers and either an issue number or title; also, all documents containing one of the following words in the source string (journal, J., conference, conf., proceeding, proc., symposium, meeting, colloquium, annual, congress).
- Source/no title: documents not in one of the previous types that had a source string, but no title string.
- Not classified: all remaining documents.

Table 5 shows the numbers of documents and average citation counts to those documents for each of the four document groups used in Table 4. Inclusion of books is clearly important given that they are cited much more highly on average than articles in all four document groups. Handbooks have mixed properties. For non-source and older documents, they are cited slightly more on average than articles. However, those handbook documents from the Src9611 group are cited less often than articles. Inspection of some of the items in this set suggests that most handbook source (as opposed to non-source) items are individual chapters from handbooks rather than full handbooks, and that these lower citation rates may be due to

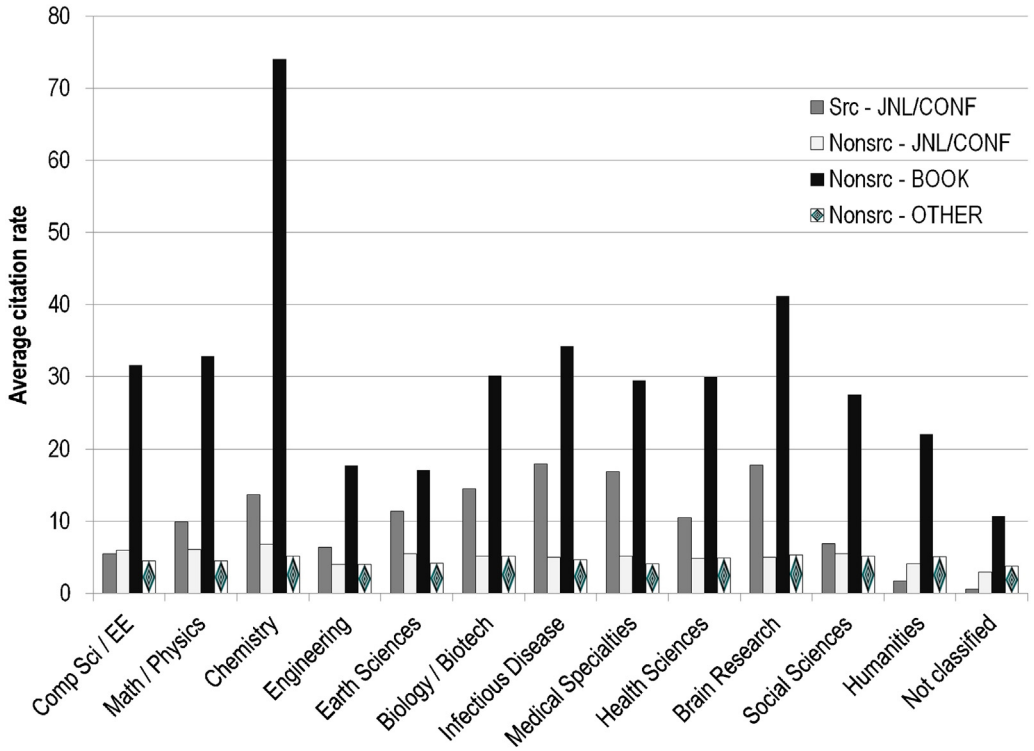

Fig. 2. Average citation counts to documents (1996–2011) in the NS map by major field and document type.

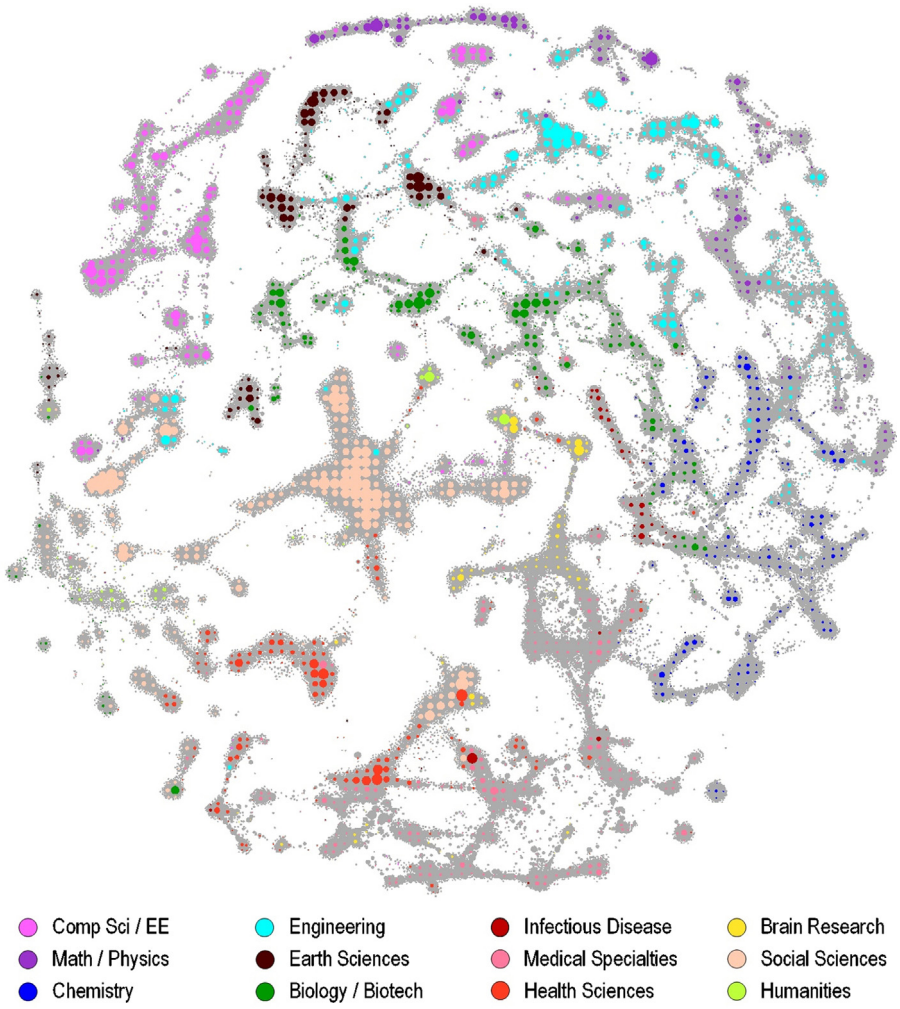

Fig. 3. Number of books by location on the NS map. Dot sizes (areas) reflect the number of books in each sector of the map.

this higher level of differentiation. The 'Source/no title' and 'Not classified' types have very similar citation characteristics, suggesting that no separation of these two types was needed.

The fact that books are more highly cited on average than journal articles holds by major field as well. Average citation rates for documents published from 1996 to 2011 in groups Src9611 and Nonsrc9611 are compared by type in Fig. 2. Source-indexed journal and conference articles are more highly cited than non-source articles in all fields except Computer Science, where the source and non-source values are very close, and the Humanities. In the Social Sciences, source articles are cited only slightly more than non-source articles. Average citation counts to all cited non-source documents other than those tagged as articles or books are similar to those of non-source articles for nearly all major fields. We note that the non-source citation rates are somewhat inflated in this analysis because items with only one citation were not included in the map. If these had been included, average citations to non-source items would have been lower than reported here. However, given the huge difference between citation rates to books and articles, including books cited only once would not have changed the finding that books are much more highly cited on average than are articles. Thus, we agree with Kousha, Thelwall, and Rezaie (2011) who found that citation counts to books are sufficiently high to enable their evaluation by citation analysis.

Although books are relatively highly cited in all major fields of science, they are not evenly distributed across the NS map, as shown in Fig. 3. The Social Sciences has the largest number of books; nearly 3% of all 1996–2011 documents in the Social Sciences part of the map are books. Engineering, Computer Science, Earth Sciences, and the Humanities are also well represented, with books comprising from 1.4 to 1.9% of documents from 1996 to 2011. Chemistry, Medical Specialties, and Brain Research all rely far less on books, which comprise less than 0.4% of the documents in these fields.

5. Summary

This study reports the first large-scale map of science that includes cited non-source items in bulk. Inclusion of non-source items not only adds them to the map, but also significantly increases the coverage of source items as well by virtue of the fact that they cite non-source items. Although a majority of the non-source items included in the map are more than 15 years old, one third of all non-source items are from the most recent 15 years. Addition of non-source items to the map changes its structure in significant ways. The Social Sciences have a much more central position in the map when non-source items are included, and the balance between major fields is perturbed. Books have been found to be present in nearly all areas of the map, and given the fact that they are more highly cited than articles, many of them undoubtedly play the role of aggregators in the direct citation structure. The role of highly cited books in topic formation and in our perceptions of the structure of science are research topics that deserve future attention. Taken together, these observations suggest that the current definition of source materials by the primary citation databases is not sufficient to fully and accurately characterize the current structure of science. Addition of books to these citation databases (Thomson's Book Citation Index, and the Scopus Book Titles Expansion program) will certainly help in this regard. However, since non-source journal articles and conference papers significantly outnumber books (see Table 5, Nonsrc9611), expansion of journal and conference source materials would likely have an even greater effect.

Nederhof et al. (2010) opined that limiting to source materials might "offer an incomplete view on scholarly citation impact in (1) fields in which journals are not of prime importance as means of scholarly communication and/or (2) fields in which important journals are covered poorly by WoS." We would go further. Given that non-source items are found in significant numbers in all areas of the map of science, we suggest that they should be included in all maps and citation analyses to the extent possible.

To the best of our knowledge, this map is also the largest map of science ever created, containing over 43 million documents across all of the sciences. Larger and more accurate maps will undoubtedly be created in the future. We suggest that maps such as these could be the basis for new classification systems that will allow accurate classification of non-source items alongside source items. Not only will such classification systems be useful in understanding the impact of non-source items on the structure of science, but they could be the basis for inclusion of non-source documents in research evaluation studies and exercises as well.

Acknowledgements

We are indebted to Ludo Waltman and Nees Jan van Eck of CWTS for making their clustering code available for general use, and to our colleague Michael Patek who performed the data extraction and code runs that made this map possible.

References

Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M., et al. (2009). Clickstream data yields high-resolution maps of science. *PLoS ONE*, *4*(3), e4803.

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., et al. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE*, *7*(7), e39464.

Boyack, K. W., & Klavans, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology*, *65*(4), 670–685. http://dx.doi.org/10.1002/asi.22990

Butler, L., & Visser, M. S. (2006). Extending citation analysis to non-source items. *Scientometrics*, *66*(2), 327–343.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377.

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology*, *61*(7), 1386–1409.

Chen, C., & Kuljis, J. (2003). The rising landscape: A visual exploration of superstring revolutions in physics. *Journal of the American Society for Information Science and Technology*, *54*(5), 435–446.

Chi, P.-S. (2013). Do non-source items make a difference in the social sciences? In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. Moed (Eds.), *14th international conference of the international society for scientometrics and informetrics* Vienna, Austria, (pp. 612–625).

Franceschini, F., & Maisano, D. (2011). Structured evaluation of the scientific output of academic research groups by recent h-based indicators. *Journal of Informetrics*, *5*, 64–74.

Garfield, E. (1973). Historiographs, librarianship, and the history of science. In C. H. Rawski (Ed.), *Toward a theory of librarianship: Papers in honor of Jesse Hauk Shera* (pp. 380–402). Metuchen, NJ: Scarecrow Press.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, *44*(2), 193–215.

Hicks, D. (2004). The four literatures of the social science. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems* (pp. 473–495). Dordrecht: Springer.

Huang, M., & Chang, Y. (2008). Characteristics of research output in social sciences and humanities: From a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, *59*(11), 1819–1828.

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, *62*(11), 2147–2164.

Leydesdorff, L. (2003). Can networks of journal–journal citations be used as indicators of change in the social sciences? *Journal of Documentation*, *59*, 84–104.

Leydesdorff, L. (2008). Caveats for the use of citation indicators in research and journal evaluations. *Journal of the American Society for Information Science and Technology*, *59*(2), 278–287.

Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). OpenOrd: An open-source toolbox for large graph layout. *Proceedings of SPIE – The International Society for Optical Engineering*, *7868*, 786806.

Moed, H. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and humanities: A review. *Scientometrics*, *66*(1), 81–100.

Nederhof, A. J., van Leeuwen, T. N., & van Raan, A. F. J. (2010). Highly cited non-journal publications in political science, economics and psychology: A first exploration. *Scientometrics*, *83*, 363–374.
Noyons, E. C. M., & Calero-Medina, C. (2009). Applying bibliometrics mapping in a high level science policy context: Mapping the research areas of three Dutch Universities of Technology. *Scientometrics*, *79*(2), 261–275.
Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*, 265–269.
Tijssen, R. J. W., & van Leeuwen, T. N. (1995). On generalising scientometric journal mapping beyond ISI's journal and citation databases. *Scientometrics*, *33*(1), 93–116.
van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.
van Leeuwen, T. N. (2006). The application of bibliometric analysis in the evaluation of social science research: Who benefits from it, and why it is still feasible. *Scientometrics*, *66*(1), 133–154.
Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392.
White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, *32*(3), 163–171.
Zuccala, A., & Guns, R. (2013). Comparing book citations in humanities journals to library holdings: Scholarly use versus perceived cultural benefit. In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. Moed (Eds.), *14th international conference of the international society for scientometrics and informetrics* Vienna, Austria, (pp. 353–360).